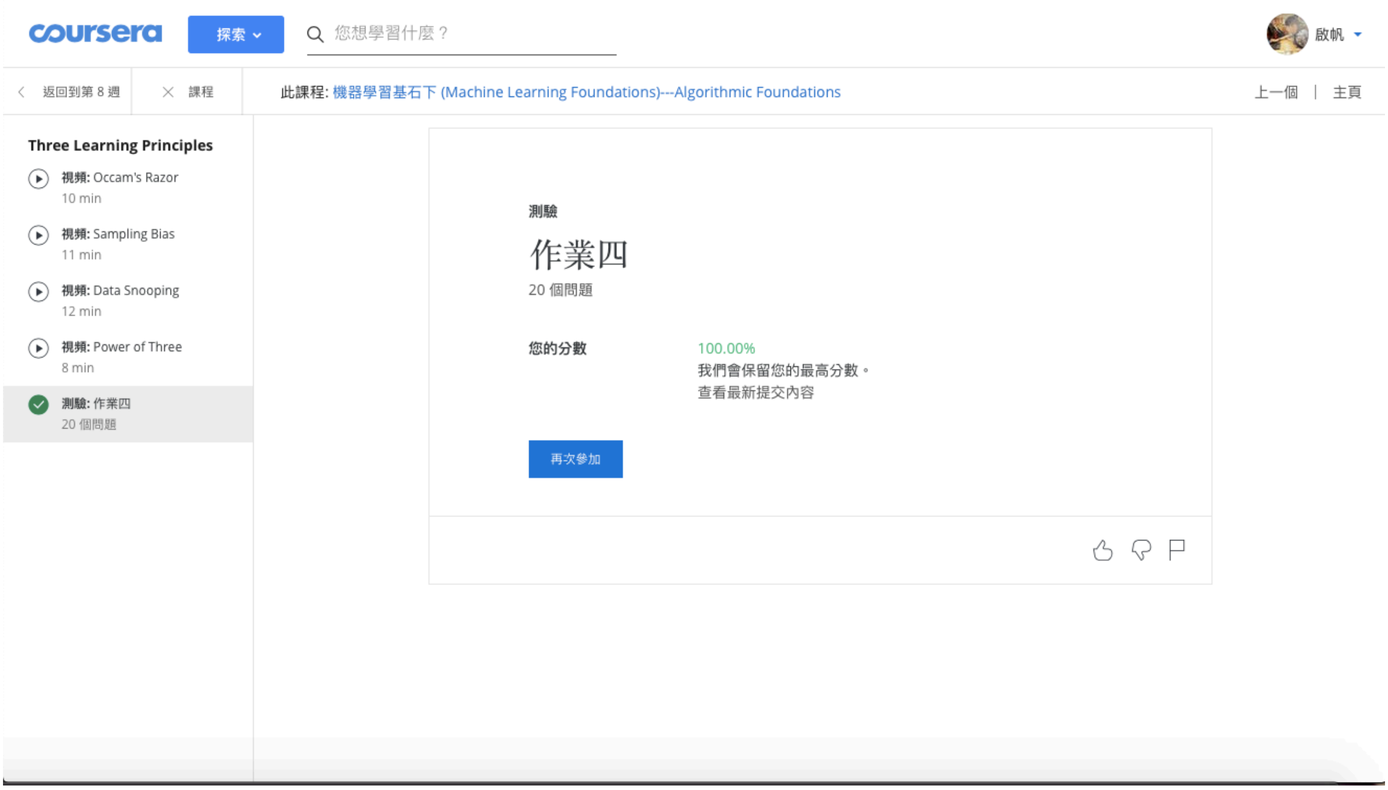
1.



2.

$$\nabla E_{aug}(\boldsymbol{w}) = \nabla E_{in}(\boldsymbol{w}) + \nabla\left(\frac{\lambda}{N}\boldsymbol{w}^T\boldsymbol{w}\right)$$

$$= \nabla E_{in}(\boldsymbol{w}) + \frac{2\lambda}{N}\boldsymbol{w}$$

So, the update rule is:

$$\boldsymbol{w}(t+1) \leftarrow \boldsymbol{w}(t) - \eta\nabla E_{aug}\big(\boldsymbol{w}(t)\big) = \boldsymbol{w}(t) - \eta\left(\nabla E_{in}\big(\boldsymbol{w}(t)\big) + \frac{2\lambda}{N}\boldsymbol{w}(t)\right)$$

$$= \left(1 - \frac{2\eta\lambda}{N}\right)\boldsymbol{w}(t) - \eta\nabla E_{in}\big(\boldsymbol{w}(t)\big)$$

3.

$$\text{Let } A = (-1,0), B = (\rho,1), C = (1,0)$$

$$E_{loo} = \frac{1}{3}\big(error(leave\ B\ out) + error(leave\ A\ out) + error(leave\ C\ out)\big)$$

Let h_{xy} be the corresponding linear hypothesis which have the lowest square error on point x and y, which is the line constructed by x and y (the square error is zero).

$$h_{AC}(x) = 0, error(leave\ B\ out) = 1$$

$$h_{BC}(x) = \frac{1}{\rho-1}(x-1), error(leave\ A\ out) = \big(h_{BC}(-1) - 0\big)^2 = \frac{4}{(\rho-1)^2}$$

$$h_{AB}(x) = \frac{1}{\rho+1}(x+1), error(leave\ C\ out) = \big(h_{AB}(1) - 0\big)^2 = \frac{4}{(\rho+1)^2}$$

$$\text{So, } E_{loo}(\rho) = \tfrac{1}{3}\left(1 + \tfrac{4}{(\rho-1)^2} + \tfrac{4}{(\rho+1)^2}\right)$$

4.

$$Let\ E_{aug}(\boldsymbol{w}) = \frac{\lambda}{N}\|\boldsymbol{w}\|^2 + E_{in}(\boldsymbol{w}) = \frac{\lambda}{N}\|\boldsymbol{w}\|^2 + \frac{1}{N}\|\boldsymbol{Xw} - \boldsymbol{y}\|^2$$

$$\nabla E_{aug}(\boldsymbol{w}) = \frac{2\lambda}{N}\boldsymbol{w} + \nabla E_{in}(\boldsymbol{w}) = \frac{2\lambda}{N}\boldsymbol{w} + \frac{2}{N}(\boldsymbol{X^T X w} - \boldsymbol{X^T y})$$

SGD: $\nabla E_{aug}(\boldsymbol{w}) = \frac{2\lambda}{N}\boldsymbol{w} + \nabla E_{in}(\boldsymbol{w}) \approx \frac{2\lambda}{N}\boldsymbol{w} + 2(\boldsymbol{w^T x_n x_n} - y_n \boldsymbol{x_n})$

(use point $(\boldsymbol{x_n}, y_n)$ to approximate the true gradient)

So the update rule is:

$$\boldsymbol{w_{t+1}} = \boldsymbol{w_t} - \eta \nabla E_{aug}(\boldsymbol{w_t})$$

$$= \boldsymbol{w_t} - \eta\left(\frac{2\lambda}{N}\boldsymbol{w_t} + 2(\boldsymbol{w_t^T x_n x_n} - y_n \boldsymbol{x_n})\right)$$

$$= \left(1 - \frac{2\eta\lambda}{N}\right)\boldsymbol{w_t} - \eta\left(2(\boldsymbol{w_t^T x_n x_n} - y_n \boldsymbol{x_n})\right)$$

The update rule in Q3 looks like this:

$$\boldsymbol{w_{t+1}} = \left(1 - \frac{2\eta\lambda}{N}\right)\boldsymbol{w_t} - \eta \nabla E_{in}(\boldsymbol{w_t})$$

We can observe that $\left(1 - \frac{2\eta\lambda}{N}\right)\boldsymbol{w_t} - \eta\left(2(\boldsymbol{w_t^T x_n x_n} - y_n \boldsymbol{x_n})\right) \approx \left(1 - \frac{2\eta\lambda}{N}\right)\boldsymbol{w_t} - \eta \nabla E_{in}(\boldsymbol{w_t})$

So, the two update rules are probably approximately same, meaning that when we add virtual example $\widetilde{\boldsymbol{X}} = \sqrt{\lambda}\boldsymbol{I}, \breve{\boldsymbol{y}} = \boldsymbol{0}$ in the training data set and do normal SGD without regularization, it can reach the same result using GD with regularization.

5.

For target function $\sin(ax), x \in [0, 2\pi]$, the squared error for $h(x) = wx$ is:

$$err(w) = \int_0^{2\pi} (\sin(ax) - wx)^2 dx = -\frac{2wsin(2\pi a)}{a^2} - \frac{\cos(2\pi a)\,(\sin(2\pi a) - 8\pi w)}{2a} + \frac{8\pi^3 w^2}{3} + \pi$$

$$\frac{\partial err(w)}{\partial w} = -\frac{2sin(2\pi a)}{a^2} + \frac{\cos(2\pi a)\,(8\pi)}{2a} + \frac{16\pi^3 w}{3}$$

To solve $\min_w err(w)$, we need to solve the equation $\frac{\partial err(w)}{\partial w} = 0$. After calculation, we get:

$$w = \frac{\frac{2sin(2\pi a)}{a^2} - \frac{\cos(2\pi a)\,(8\pi)}{2a}}{\frac{16\pi^3}{3}} = \frac{3sin(2\pi a) - 6\pi a\cos\,(2\pi a)}{8\pi^3 a^2}$$

So, for each x, the level of deterministic noise is $\left|\sin(ax) - \frac{3sin(2\pi a) - 6\pi a\cos\,(2\pi a)}{8\pi^3 a^2}x\right|$